Histopathology 2006, **48**, 655–662. DOI: 10.1111/j.1365-2559.2006.02394.x

A study of Gleason score interpretation in different groups of UK pathologists; techniques for improving reproducibility

D F R Griffiths, J Melia,[1] L J McWilliam,[2] R Y Ball,[3] K Grigor,[4] P Harnden,[5] M Jarmulowicz,[6] R Montironi,[7] R Moseley,[8] M Waller,[1] S Moss[1] & M C Parkinson[9] in collaboration with the participating pathologists listed*

Department of Pathology, Cardiff University, Cardiff, [1]Cancer Screening Evaluation Unit, Institute of Cancer Research, Sutton, [2]Department of Histopathology, Manchester Royal Infirmary, Manchester, [3]Department of Histopathology, Norfolk & Norwich University Hospital, Norwich, [4]Department of Pathology, Western General Hospital, Edinburgh, [5]Department of Histopathology, St James Hospital, Leeds and [6]present affiliation: Bostwick Laboratories, London, UK, [7]present affiliation: Institute of Pathological Anatomy and Histopathology, Polytechnic University of the Marche Region (Ancona), School of Medicine, Ancona, Italy, [8]Department of Histopathology, Addenbrooke's Hospital, Cambridge and [9]University College Hospitals and Institute of Urology, UCL, London, UK

Date of submission 16 November 2005
Accepted for publication 3 January 2006

Griffiths D F R, Melia J, McWilliam L J, Ball R Y, Grigor K, Harnden P, Jarmulowicz M, Montironi R, Moseley R, Waller M, Moss S & Parkinson M C

(2006) *Histopathology* **48**, 655–662

A study of Gleason score interpretation in different groups of UK pathologists; techniques for improving reproducibility

Aims: To test the effectiveness of a teaching resource (a decision tree with diagnostic criteria based on published literature) in improving the proficiency of Gleason grading of prostatic cancer by general pathologists.

Methods: A decision tree with diagnostic criteria was developed by a panel of urological pathologists during a reproducibility study. Twenty-four general histopathologists tested this teaching resource. Twenty slides were selected to include a range of Gleason score groups 2–4, 5–6, 7 and 8–10. Interobserver agreement was studied before and after a presentation of the decision tree and criteria. The results were compared with those of the panel of urological pathologists.

Results: Before the teaching session, 83% of readings agreed within ± 1 of the panel's consensus scores. Interobserver agreement was low ($\kappa = 0.33$) compared with that for the panel ($\kappa = 0.62$). After the presentation, 90% of readings agreed within ± 1 of the panel's consensus scores and interobserver agreement amongst the pathologists increased to $\kappa = 0.41$. Most improvement in agreement was seen for the Gleason score group 5–6.

Conclusions: The lower level of agreement among general pathologists highlights the need to improve observer reproducibility. Improvement associated with a single training session is likely to be limited. Additional strategies include external quality assurance and second opinion within cancer networks.

Keywords: biopsy, education, grade, prostate carcinoma, reproducibility

*List of participating consultant pathologists in Cardiff: A. M. Rashid, N. Williams, S. Polacarz, M. Lord, S. Banarjee, G. T. Smith, E. Wessels, J. K. Murphy, C. Champ, J. Shannon, R. B. Denholm, S. Kiberu, C. G. B. Simpson, G. R. Melville Jones, B. Charnley, D. Stock, G. Evans, M. Cotter, N. Nind, M. Hayes, R. Kellett, R. Williams, S. Howell, L. Murray.

List of participating consultant pathologists in Cambridge: K. J. Arulambalam, D. M. Bailey, N. M. S. Bajallan, S. Boyle, R. A. Carr, A. Chandra, M. E. Chappell, S. B. Coghill, D. R. Davies, S. Jader, P. R. Maheswaran, C. S. F. McCormick, T. McCulloch, A. J. Molyneux, P. J. O'Donnell, P. S. Ong, A. Patterson, D. S. Peat, H. Rees, D. S. C. Rose, H. Shaikh, R. E. Smith, R. W. Stirling, L. N. Temple, J. M. Theaker, P. Thebe, S. Thomas, G. J. Tidsley, P. A. Trott, A. Verghese, P. Q. Wolfe, M. P. A. Young.

Address for correspondence: Dr David F R Griffiths MB BCh, MRCP(UK), FRCPath, Department of Pathology, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK. e-mail: griffithsdfr@cf.ac.uk

Introduction

There has been an increase in prostatic needle core biopsies aimed at the detection of gland-confined, potentially curable cancer and in the use of the widely recommended Gleason system for grading.[1–3] Information as to how this Gleason system is being applied in the UK by general histopathologists is very limited;[4] nor has the effect of different teaching modalities been assessed. Studies from the USA have shown that general pathologists have a lower level of reproducibility when using the Gleason system[5] than pathologists with a major commitment to urological pathology (uropathologists).[6] Other studies from outside the UK have demonstrated some improvement in reproducibility after various teaching strategies[7–10] and over time.[11]

Criteria for the Gleason score[12] were applied by a group of histopathologists with a special interest in urological pathology to study observer variation. (To avoid confusion with grade and difficulty in interpreting a single figure as representing one or two patterns, we have used the terms Gleason score for the sum score and Gleason pattern for morphological types 1–5.) Overall, good interobserver agreement was achieved ($\kappa = 0.62$). A decision tree was developed to assist the application of the Gleason score criteria. The purpose of the present study was to investigate the level of agreement in colleagues with a wider diagnostic interest and to assess whether the use of the decision tree would change their level of agreement.

Methods

CRITERIA AND DECISION TREE

The criteria for each Gleason pattern based on published literature (Table 1)[13,14] were agreed by a panel of nine consultants with a special interest in uropathology. Eight of the panel came from different locations in the UK and one consultant was a European representative based on his established expertise in prostate pathology and morphometry. A decision tree was developed which placed the criteria in context and provided reference images (Figure 1).

STUDY PATHOLOGISTS

A total of 24 consultant pathologists from Wales who attend joint pathology meetings took part in a half-day meeting in Cardiff to evaluate the effect of using the criteria and decision tree on observer variation of the Gleason score. In the first 1-h microscopy session, the pathologists assessed the slides as they would in normal practice with the Gleason diagram available for reference, as required. The pathologists rotated around individual microscopes; there was no conferring. After the first session, they all attended a 40-min lecture at

Table 1. Criteria

Pattern 1	Space between malignant acini < 1 acinar dimension (of the largest acinus present), margins sharp circumscribed
Pattern 2	Space between malignant acini < 1 acinar dimension (of the largest acinus present), margins lack circumscription
Pattern 3	If any of the following: Benign ducts/acini present between malignant acini
	Space between malignant acini > 1 acinar dimension (of the largest acinus present)
	2 × variation in dimension between malignant acinar dimension
	Small cribriform areas with smooth edges
	Malignant acini angulated
Pattern 4	Fusion, involving large areas or smaller groups of acini fused in a grape-like pattern, length of fusion > 4 × longer than dimension of one of the fused acini
Pattern 5	Sheets of cells including those with small lumens which are present in < 50% of the area; any commedo (tumour) necrosis; single cell invasion; signet ring cells
Dimension	Formerly width or diameter, is the narrowest part of the acinus. May refer to the largest or smallest acinus present

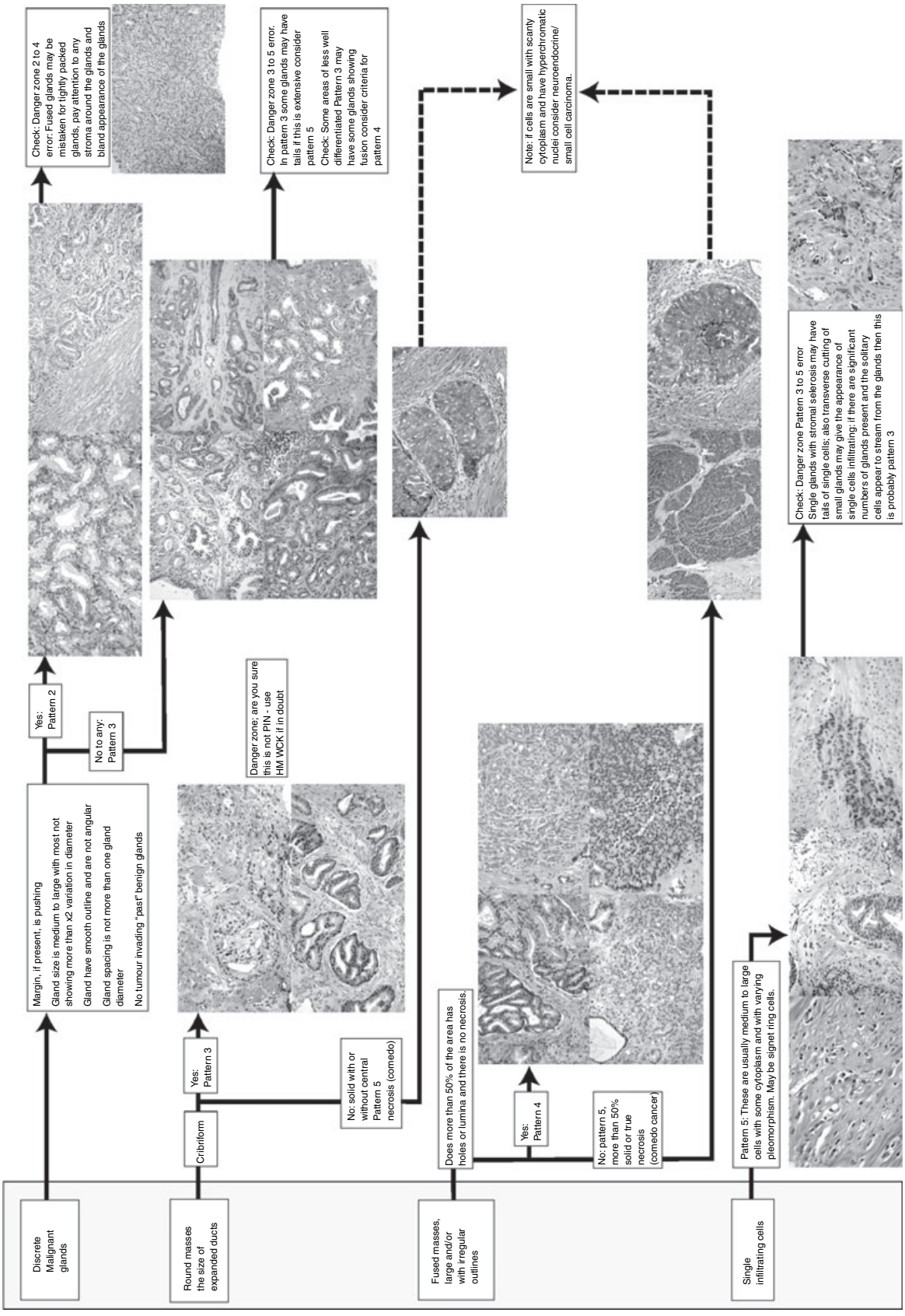

Figure 1. Diagnostic decision tree developed for Gleason grading. Colour images are provided as examples of each grade as a reference. This figure is reduced from the original colour A3 (420 × 297 mm) sized graphic. Some images have been substituted or rescaled as a result of user feedback.

which the study criteria were illustrated and the use of the decision tree (Figure 1) was explained. No photomicrographs from test cases were shown; there was opportunity to ask questions and receive further clarification. After a short break, during which the slides were renumbered, pathologists then re-read the same slides in a different order in a second 1-h microscopy session with the decision tree and criteria available for reference.

SLIDES AND READINGS

In the study by the uropathology panel a set of slides was assembled by asking each pathologist to contribute 10 haematoxylin and eosin-stained sections from prostatic needle core biopsies from consecutive clinical series within the score groups 2–4, 5–6, 7 and 8–10.[12] Most were from the year 2000, but some laboratories had to go back to 1996 to find slides with Gleason score 2–4. A random sample of slides was selected within each score group and 20 slides were read in each of six circulations. Some slides were re-read to assess intra-observer variation, resulting in a total of 81 slides read at least once. From these slides, 20 were selected for the Cardiff evaluation study stratified by the Gleason scores 2–4, 5–6, 7 and 8–10 originally assigned in each urological pathologist's laboratory. This resulted in the inclusion of five slides within each of the four Gleason groups. No slides were masked to show only selected areas.

DATA COLLECTION

Each study pathologist taking part in the evaluation study completed a questionnaire to collect data on their workload, training and age. At the meeting, the pathologists assessed the slides as they would in clinical practice, with the availability of the Gleason diagram. They used a standardized proforma to record the major, minor and tertiary patterns. The questionnaires and proformas were sent to the Cancer Screening Evaluation Unit, where the data were anonymized before data entry and analysis.

STATISTICS

Determination of panel consensus values

Some comparisons were to be made between the readings of the urological panel and those of the study pathologists. For the Gleason score, the panel consensus was first calculated separately for the major and minor patterns for each slide by taking the median. The panel consensus score was the sum of these values. The scores were grouped as 2–4, 5–6, 7 and 8–10. Using the distribution of panel's score values, the slides were classified as easy or difficult: easy slides (11/20) were those where all nine members reported the same score group, or eight of the panel recorded the same and the other member reported an adjacent category. All other slides (9/20) were considered difficult.

Simple descriptive data

The slides were grouped according to their panel consensus score groups 2–4, 5–6, 7 and 8–10 and for each set of slides the readings of the panel and the study pathologists were studied in cross-tabulations.

Kappa

The interobserver agreement for each variable was studied using kappa statistics (κ) to assess the measurement of overall agreement adjusted for the agreement expected by chance.[15,16] Values from 0 to 1 indicate some level of agreement exceeding chance. It is suggested that κ values of 0.00–0.20 reflect slight agreement, values of 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement and ≥ 0.81 almost perfect agreement.[16] Negative values of κ represent systematic disagreement. Kappa statistics where there are more than two observers and more than two categories of the variable being analysed were used.[17] An overall κ was calculated for all pathologists across all patterns. Kappa was also calculated for all pathologists using slides grouped according to their consensus score, 2–10, and consensus Gleason score groups 2–4, 5–6, 7 and 8–10. Intra-observer variation was studied by calculating κ statistics for each pathologist using their readings before and after the presentation of the criteria.

Relationship of demographic and outcome variables

The relationship of personal characteristics was studied for two outcome variables: interobserver agreement with the panel and the proportion of readings recorded by the study pathologists as Gleason score 2–4. The distribution of each characteristic was assessed and they were each divided into two categories using the median values. The relationship with κ for interobserver agreement with the panel was studied in simple tabulations. The relationship with the reporting of Gleason scores 2–4 was studied in logistic regression analyses.

Results

DEMOGRAPHICS OF THE STUDY PATHOLOGISTS

Their mean age was 47 years (SD 8.3) and they had held the MRCPath for a mean of 14 years (SD 7.7).

Table 2. Comparison of the consensus score of the uropathology panel with the readings recorded by the study pathologists before the presentation

	Readings of Cardiff study pathologists									
Panel consensus	−4	−3	−2	−1	0 (%)	+1	+2	+3	+4	Total
6	1	3	15	24	85 (51)	22	9	3	5	167
7	4	8	22	50	92 (48)	15	1	0	0	192
8	0	2	1	18	92 (48)	16	0	0	0	48
9	0	0	1	9	8 (33)	6	0	0	0	24
10	0	0	6	19	23 (48)	0	0	0	0	48
Total	5	13	45	120	219 (46)	59	10	3	5	479

Most of them (88%) reported that less than 20% of their workload was uropathology and the mean number of prostatic specimens read per year was 208.6 (SD 150.5). They had used the Gleason grading for a mean of 10.1 years (SD 5.6) and most had learned about the grading (21/24) from books and journals.

OBSERVER AGREEMENT

Before the criteria and decision tree were presented to the pathologists, 46% of the readings of the individual Gleason scores 2–10 recorded by the pathologists agreed exactly with the panel's consensus scores (Table 2). More of the pathologists' readings were below (38%) than above (16%) the consensus score. The study pathologists recorded 32 readings of Gleason score 2–4, while none of the slides had a panel consensus Gleason score of 2–4 (although the original scores of five slides recorded in the laboratories were within score groups including 2–4).

After the criteria and decision tree were presented to the pathologists, 52% of the readings of the individual Gleason scores 2–10 recorded by the pathologists agreed exactly with the panel's consensus scores. The effect of using the criteria and decision tree on interobserver agreement among the study pathologists was studied by analysing κ values overall and for the Gleason score groups 2–4, 5–6, 7 and 8–10 before and after the presentation (Table 3). The overall κ increased from 0.33 to 0.41. This was mainly the result of increased agreement in Gleason score groups 5–6 and 7. However, the interobserver agreement among the study pathologists was not as high as that among the members of the panel (overall κ 0.62).

These analyses were repeated separately for easy slides and more difficult slides. For the easy slides, the overall κ values for the study pathologists were 0.40 and 0.45 before and after the presentation, respectively, and 0.87 for the panel. For the more difficult slides, the overall κ values for the study pathologists were 0.15 and 0.22 before and after the presentation, respectively, and 0.18 for the panel.

Table 3. Kappas for interobserver agreement among the study pathologists before and after the presentations, and among the panel of uropathology specialists

Gleason score (submitted with slide)	Before the presentation	After the presentation	Panel
2–4	0.07	0.06	–
5–6	0.27	0.40	0.57
7	0.21	0.29	0.53
8–10	0.62	0.61	0.80
Total	0.33	0.41	0.62

Finally, the study pathologists' readings were compared with the consensus scores agreed by the panel. There was little change in the overall level of agreement between the study pathologists' readings and the panel's consensus before and after the presentation (Table 4), but improvement in agreement was seen for slides with consensus Gleason score of 5–6 (the percentage of readings agreeing with the consensus increased from 65% to 77%). The proportion of study pathologists' readings recording Gleason score group 2–4 decreased from 6% to 3%.

TERTIARY GRADE

In the first session, the tertiary grade was reported in 7% of all readings (34/479) and in 6% (28/479) the tertiary grade was higher than either the major or minor grades. These proportions are lower than those

	Participants' score (readings agreeing with the panel score are in bold)									
	2–4		5–6		7		8–10		Total	
Panel score	B	A	B	A	B	A	B	A	B	A
5–6	19	11	**109**	**127**	22	15	17	13	167	166
7	12	3	72	87	**92**	**86**	16	16	192	192
8–10	0	0	3	2	19	26	**98**	**92**	120	120
Total	31	14	184	216	133	127	131	121	479	478
Observed agreement			B 62% κ 0.45 A 64% κ 0.46							

Table 4. Agreement between study pathologists and the panel of uropathology specialists before (B) and after (A) the presentation of the criteria and decision tree

recorded by the panel for the same set of slides (16% and 9%, respectively). Among the study pathologists, the levels of agreement for the presence of a tertiary grade ($\kappa = 0.03$) and the individual grade values (3, 4 and 5: all < 0.11) were very low compared with those for the panel ($\kappa = 0.49$ and $\kappa = 0.52$ to $\kappa = 0.48$, respectively). There was no improvement in the recording of a tertiary grade by the study pathologists at the second session.

INTRA-OBSERVER AGREEMENT

Intra-observer agreement was analysed for each study pathologist by comparing readings before and after the presentation. Kappas ranged from 0.07 to 0.82. The distribution of the κ according to the categories of slight, fair, moderate, substantial and almost perfect agreement was two (8%), nine (38%), six (25%), six (25%), and one (4%), respectively. Low κ indicates marked change in scoring by pathologists, whereas high κ indicates little change in scoring after the presentation.

RELATION OF STUDY PATHOLOGISTS' REPORTING TO DEMOGRAPHIC FACTORS

Before the presentation, the study pathologists had more agreement with the panel if they were relatively young (≤ 47 years $\kappa = 0.51$, > 47 years $\kappa = 0.41$), had held the MRCPath for fewer years (≤ 14 years $\kappa = 0.52$, > 14 years $\kappa = 0.39$), saw more prostatic specimens per year (≤ 180 cases $\kappa = 0.37$, > 180 cases $\kappa = 0.52$), or had learned about the Gleason score on a training course (yes $\kappa = 0.58$, no $\kappa = 0.40$). After the presentation, the effect of these characteristics tended to disappear.

Before the presentation, the study pathologists were more likely to report Gleason scores 2–4 if they were older [odds ratio (OR) 4.73 per 10 years' increase $P < 0.001$] or had not learned to use the Gleason score on a training course (OR 0.15 associated with course compared with no course $P < 0.01$). In a multiple regression analysis when both factors were considered together, only age remained statistically significant (OR 3.94 for 10 years' increase in age $P < 0.001$). As for overall interobserver agreement, the effects of personal characteristics seemed to disappear after training.

Discussion

This study has two main findings: first, it has demonstrated that the agreement on Gleason score is not as good among a group of pathologists with a wider diagnostic interest (generalists) as it is within a panel of uropathologists; second, it has shown that a single brief structured educational exercise, while improving some aspects of agreement with the uropathologists, has little overall impact on agreement. These findings have important implications for patient management and pathologists' education.

The study pathologists did not show as good agreement among themselves as the uropathologists, or with a uropathologist's consensus diagnosis, as concluded in two studies from the USA.[5,11] The present study agrees with American studies, which have shown that the level of agreement among pathologists is significantly lower if they diagnose fewer cases per year,[8] are older[7] or have not learned about the Gleason scores from a course or meeting.[5] Overall, the level of agreement for the same set of slides decreased with the level of interest and expertise in uropathology, as shown in both the US studies[5,6] and the present study.

The present Cardiff study also agrees with the observation that generalists more frequently underscore than overscore.[5,7,8,11] The methods of calculating

the proportion of readings underscored varied between studies: for slides with Gleason score 7 proportions of 47%[5] and 55%[11] have been reported, with Gleason score 5–7 a proportion of 36%[8] and overall 23%.[7] Underscoring by generalists occurs more frequently in older than younger pathologists.[7] Further evidence of the difference in level of agreement between general histopathologists and our panel of histopathologists with a special interest in urological pathology comes from a pilot study which we conducted among 31 histopathologists who had a wide range of interest in uropathology attending a clinical cytology meeting in Cambridge. Only 54% of the readings for the individual Gleason sum scores 2–10 recorded by the Cambridge study pathologists agreed exactly with the panel's consensus scores and, similar to the current study, the Cambridge study pathologists tended to underscore the slides. Further results are available from the corresponding author.

The finding of underscoring by generalists concurs with experience in referral practice.[18] However, these comparisons should be treated with caution because of differences in methodology, such as the selection of slides and differing pathologist experience and culture. It is likely that the levels of agreement found in all studies may be higher than those which might be achieved in practice, as usually only one slide or image from each case was examined. Having all sections from a case available for examination may reduce agreement further.

The feedback from the Cardiff evaluation showed that 22 of the 24 pathologists found the meeting helpful to their diagnostic practice and 20 reported that it would change their clinical practice. Four pathologists would have liked more time to digest the new information.

Our finding of only limited improvement in performance immediately after structured teaching differs from three previous studies, which have shown substantial improvement in grading after a web-based tutorial,[9,10] a structured lecture[8] or distributed material including reference images.[8] The present study differs from these previous studies on the teaching of the Gleason system in that only needle biopsies were used as test material for scoring; others have either used images[7,9,10] or mixed needle biopsies with sections from radical prostatectomy specimens;[8] some studies have included images or biopsies with masked areas, thus potentially improving agreement.[8] Finally, certain studies already mentioned have incorporated only test material in which uropathologists were in complete agreement;[7,8] these sections must therefore have been equivalent to the present study's 'easy cases'.

The slides in this study were selected from those used by the panel of urological pathologists[12] to represent a range of easy to more difficult slides. The slides studied by the panel had been obtained from consecutive clinical series ensuring that they reflected the range of appearances seen in routine clinical practice. Finally, none of the study pathologists was naive to the Gleason system. Indeed, their agreement with the panel for easy cases was good ($\kappa = 0.57$) before the teaching, leaving only limited room for improvement with these cases.

Nevertheless, there was some improvement in agreement with the panel for slides with consensus Gleason scores of 5–6. However, this may have been achieved by simply up-grading scores 2–4. Disappointingly, there was no improvement in the clinically important consensus Gleason score 7 cases. This may reflect the panel's difficulty in separating Gleason patterns 3 and 4 in some cases,[12] suggesting that there is a particular challenge in this part of the Gleason system. Further improvement in grading by pathologists already familiar with the Gleason system may require continuing discussion with feedback on performance such as that which might be achieved in an external quality assurance scheme.

Agreement for the presence and level of a tertiary pattern was poor for both the panel and the study pathologists. Although the tertiary pattern may have prognostic value when it is higher than the major and minor patterns,[19] its use in clinical practice will be unreliable unless the accuracy of its identification can be improved.

It is important to improve the accuracy of prostatic cancer grading because it is currently a component determining the course of patient management. Approaches to consider include obtaining a second opinion in those cases where the grade could significantly influence management. This has been shown to be effective for grading of prostatic cancer.[20] Further evidence is needed on the best approach and the planned UK uropathology EQA scheme will provide important information. All of these possible aids to accuracy will have important resource and management implications for cancer networks.

Acknowledgements

We thank the Prostate Research Campaign UK for funding this research. We are very grateful to all the pathologists who took part in the study days, and to Angharad Williams who helped to organize the circulations and database. We also thank Adrian Shaw of Media Resources, Cardiff University for the decision tree graphic.

References

1. The Royal College of Pathologists. *Standards and minimum datasets for reporting common cancers. Minimum dataset for prostate cancer histopathology reports.* London: The Royal College of Pathologists 2000.
2. Boccon-Gibod L, van der Kwast TH, Montironi R, Boccon-Gibod L, Bono A. Handling and pathology reporting of prostate biopsies. *Eur. Urol.* 2004; **46;** 177–181.
3. Srigley JR, Amin MB, Humphrey PA. *Prostate gland. Protocol applies to invasive carcinomas of the prostate gland.* Based on AJCC/UICC TNM, 6th edn, January 2003. Chicago: College of American Pathologists 2003.
4. Lessells AM, Burnett RA, Howatson SR *et al.* Observer variability in the histopathological reporting of needle biopsy specimens of the prostate. *Hum. Pathol.* 1997; **28;** 646–649.
5. Allsbrook WC Jr, Mangold KA, Johnson MH *et al.* Interobserver reproducibility of Gleason grading of prostatic carcinoma: general pathologist. *Hum. Pathol.* 2001; **32;** 81–88.
6. Allsbrook WC Jr, Mangold KA, Johnson MH *et al.* Interobserver reproducibility of Gleason grading of prostatic carcinoma: urologic pathologists. *Hum. Pathol.* 2001; **32;** 74–80.
7. Egevad L. Reproducibility of Gleason grading of prostate cancer can be improved by the use of reference images. *Urology* 2001; **57;** 291–295.
8. Mikami Y, Manabe T, Epstein JI *et al.* Accuracy of Gleason grading by practicing pathologists and impact of education on improving agreement. *Hum. Pathol.* 2003; **34;** 658–665.
9. Kronz JD, Silberman MA, Allsbrook WC Jr. *et al.* Pathology residents' use of a Web-based tutorial to improve Gleason grading of prostate carcinoma on needle biopsies. *Hum. Pathol.* 2000; **31;** 1044–1050.
10. Kronz JD, Silberman MA, Allsbrook WC, Epstein JI. A web-based tutorial improves practicing pathologists' Gleason grading of images of prostate carcinoma specimens obtained by needle biopsy: validation of a new medical education paradigm. *Cancer* 2000; **89;** 1818–1823.
11. Renshaw AA, Schultz D, Cote K *et al.* Accurate Gleason grading of prostatic adenocarcinoma in prostate needle biopsies by general pathologists. *Arch. Pathol Lab. Med.* 2003; **127;** 1007–1008.
12. Melia J, Moseley R, Ball RY *et al.* A UK-based investigation of inter- and intra-observer reproducibility of Gleason grading of prostatic biopsies. *Histopathology* 2006; DOI: 10.1111/j.1365-2559.2006.02393.x
13. Gleason DF, Mellinger GT, The Veterans Administration Cooperative Urological Research Group. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. *J. Urol.* 1974; **111;** 58–64.
14. Bostwick DG, Dundore PA. Biopsy pathology of the prostate (grading needle biopsies). In London, Weinheim: Chapman & Hall Medical 1997; 141–166.
15. Cohen JA. A coefficient of agreement for nominal scales. *Educ. Psychol. Measure.* 1960; **20;** 37–46.
16. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ* 1992; **304;** 1491–1494.
17. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol. Bull.* 1971; **76;** 378–382.
18. Iczkowski KA, Bostwick DG. The pathologist as optimist: cancer grade deflation in prostatic needle biopsies. *Am. J. Surg. Pathol.* 1998; **22;** 1169–1170.
19. Pan CC, Potter SR, Partin AW, Epstein JI. The prognostic significance of tertiary Gleason patterns of higher grade in radical prostatectomy specimens: a proposal to modify the Gleason grading system. *Am. J. Surg. Pathol.* 2000; **24;** 563–569.
20. Carlson GD, Calvanese CB, Kahane H, Epstein JI. Accuracy of biopsy Gleason scores from a large unropathology laboratory: use of a diagnostic protocol to minimize observer variability. *Urology* 1998; **51;** 525–529.